Advanced Econometrics – Model with binary dependent variable

Prepared by: Daria Ivanushenko

Date: 02.06.2021

Abstract:

Aim of this paper is to identify characteristics which have impact on the income such as education, age, gender, race, marital status, occupation and etc. One of the main tool which will be used to retrieve proper results is logit / probit models for binary dependent variables in combination with correspondent test and statistical measures like Linktest, Hosmer-Lemeshow test, calculation of odds ratio and pseudo R-Squares, testing hypothesis for restricted and unrestricted models. All the mentioned test and measures should held to identify the correctness of the chosen model and its efficiency. Results of Hosmer0Lemeshow test and Linktest showed that our model has missing variables which could have important impact on income.

1. Introduction

Main aim of this paper is to identify influencing factors of the income level and way of their impact. Additionally, several crucial questions will be described as well. Topic of gender inequality is the most discussed in media and in our study we want try to understand the impact of gender and how strong it can affect the income. Moreover, Another very common statement that you can hear more often among young generation is the next: Education doesn't not play as important role as it used to in growing your income. Using information about type of education and number of years of schooling we will try to provide some evidences to support or reject the claim.

2. Literature Review

Before starting the analysis, education materials regarding econometric models with binary variables were reviewed. Additionally, in research paper "Examining the Factors Affecting Personal Income: An empirical Study Based on Survey Data in Chinese Cities" it was proved that education and marital play important role on growing income. In our paper we will analyze mainly United States and some Latin American countries and we will try to check if the same tendency applies not only for Chinese countries.

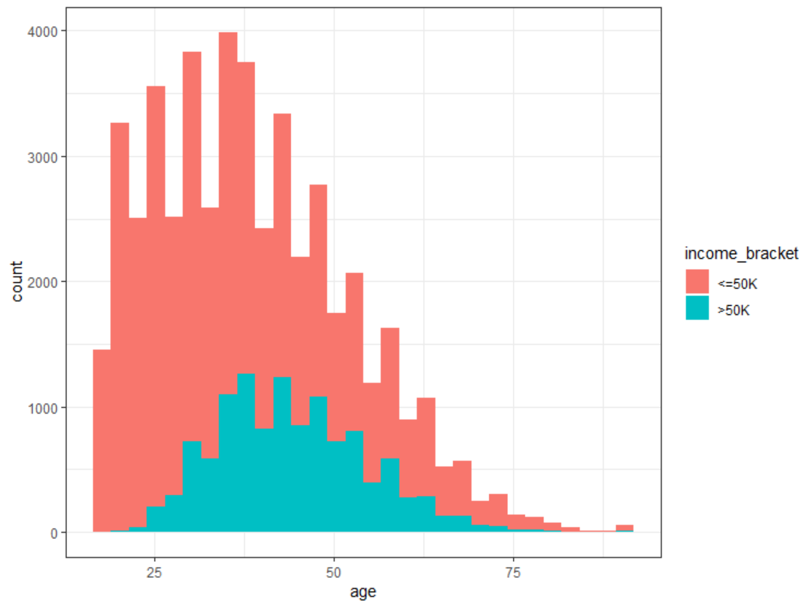
3. Data

Data used in the study was taken from publicly available resource known as Kaggle. Dataset consist of 48842 observations and 15 variables. Below you can find full description of the variables:

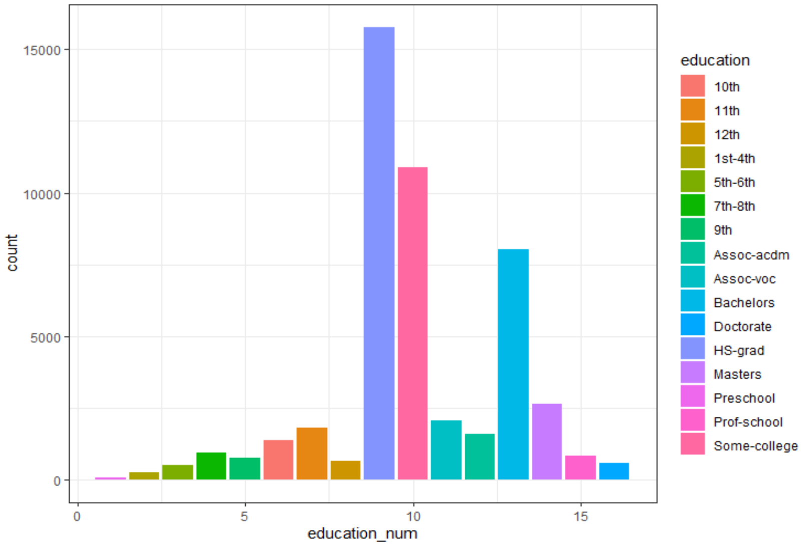
- age: continuous.
- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- fnlwgt: continuous.
- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- education-num: continuous.
- marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

- occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex: Female, Male.
- capital-gain: continuous.
- capital-loss: continuous.
- hours-per-week: continuous.
- native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

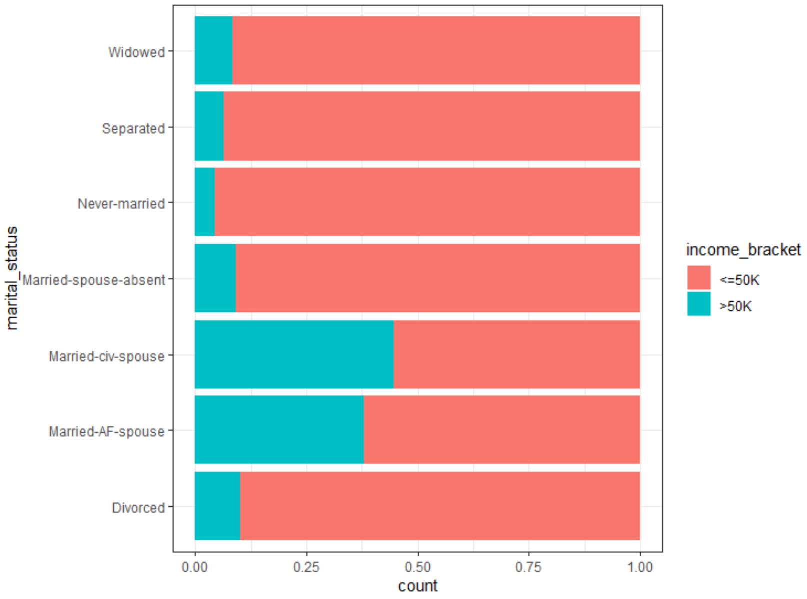
EDA:



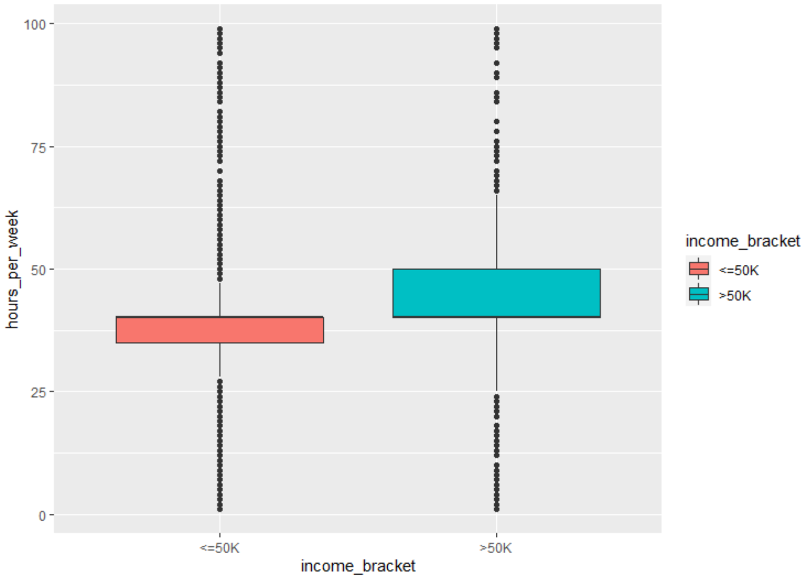
Distribution of the age variable. Distribution is right-skewed. The most frequent age is approximately between 30-40 years old. Minimum value for age is 17 and maximum value for age is 90. Group having income more than 50K is smaller comparing to group with income less than 50K. Group with 50K or more increases with age while group with less than 50K is dropping.



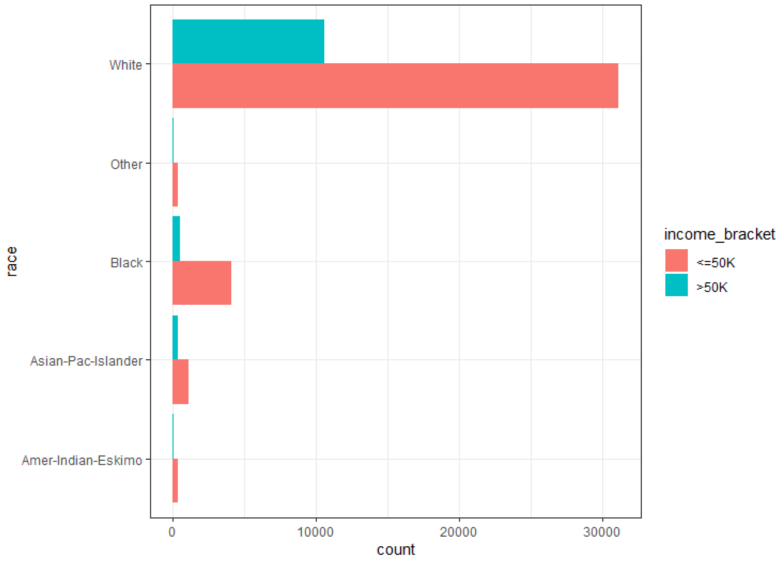
From the histogram for years of schooling variable we can notice that the most frequent value is 9 years which corresponds to high school graduation. We do not see any skewness on the graph, so we can conclude that distribution of years of schooling is rather symmetrical. The maximum value is 16 years and minimum value is 1 year of schooling.



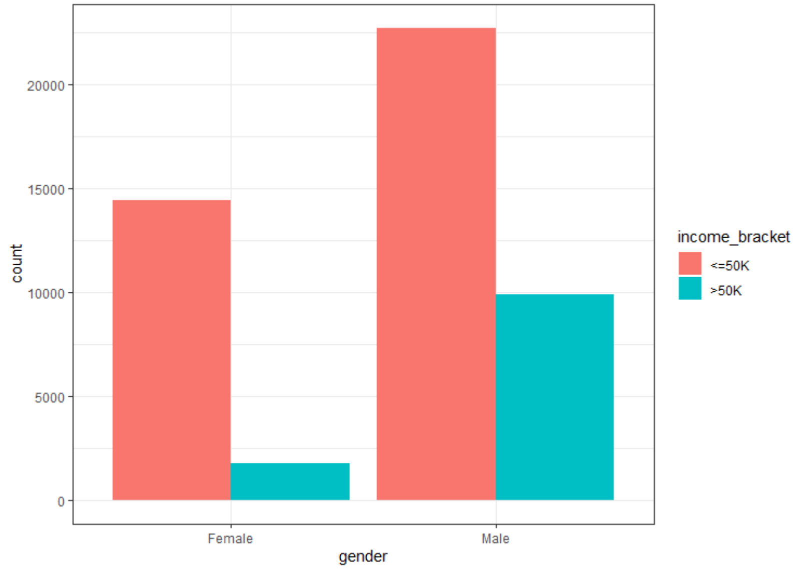
The biggest group with income equal to or more than 50K is for those that have civil marriage or have spouse in the Armed-Force. The smallest group that has income less 50K is those that are never married, separated or widowed.



On the graphed it is illustrated that group which has income 50K or more works more hours during the week comparing to group which has less than 50K.



The most representative race in our data is white where majority has income 50K or less. Other races underrepresented in our data hence they will be combined into one category.



In our data male groups is more represented than female. We have more male respondents with income higher than 50K comparing to female group.

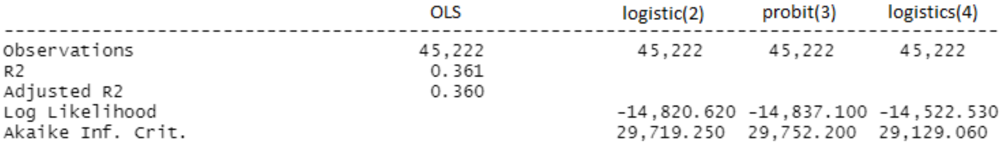
Data Preprocessing:

Variable income was recorded into 0 and 1 where 1 corresponds to group with income higher than 50K. Column final weight was removed from the data set due to the fact that it’s a weight calculated based on demographical factors already included in the dataset. Underrepresented classes were merged together. Also, it will be helpful for the interpretation.

4. Model

Due to the fact that our independent variable is binary, model with binary dependent variable will be used. Linear probability model, logit, probit models were estimated. Based on table 1 we can conclude that logistic (4) model (Logit Non-liner Model) should be chosen based on the smallest value for the Akaike information criteria which is equal to 29,129.060.

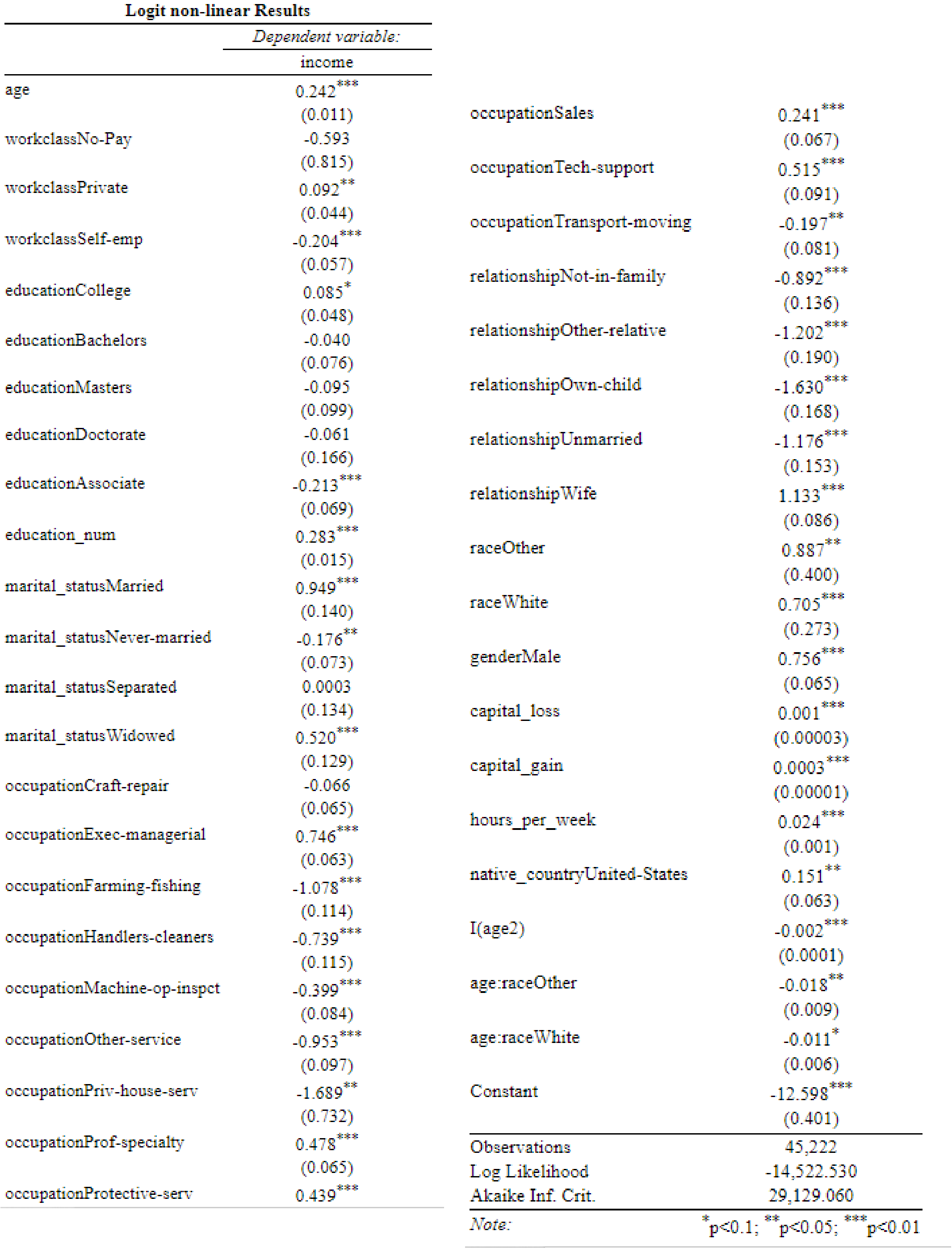
Table 1. Akaike Information Criteria



```
                                          OLS              logistic(2)    probit(3)    logistics(4)
------------------------------------------------------------------------------------------
Observations                            45,222           45,222       45,222       45,222
R2                                       0.361
Adjusted R2                              0.360
Log Likelihood                                       -14,820.620 -14,837.100 -14,522.530
Akaike Inf. Crit.                                     29,719.250  29,752.200  29,129.060
```

Following Table 2 will present chosen model.

Table 2. Models



Logit non-linear Results

	Dependent variable:
	income
age	0.242***
	(0.011)
workclassNo-Pay	-0.593
	(0.815)
workclassPrivate	0.092**
	(0.044)
workclassSelf-emp	-0.204***
	(0.057)
educationCollege	0.085*
	(0.048)
educationBachelors	-0.040
	(0.076)
educationMasters	-0.095
	(0.099)
educationDoctorate	-0.061
	(0.166)
educationAssociate	-0.213***
	(0.069)
education_num	0.283***
	(0.015)
marital_statusMarried	0.949***
	(0.140)
marital_statusNever-married	-0.176**
	(0.073)
marital_statusSeparated	0.0003
	(0.134)
marital_statusWidowed	0.520***
	(0.129)
occupationCraft-repair	-0.066
	(0.065)
occupationExec-managerial	0.746***
	(0.063)
occupationFarming-fishing	-1.078***
	(0.114)
occupationHandlers-cleaners	-0.739***
	(0.115)
occupationMachine-op-inspct	-0.399***
	(0.084)
occupationOther-service	-0.953***
	(0.097)
occupationPriv-house-serv	-1.689**
	(0.732)
occupationProf-specialty	0.478***
	(0.065)
occupationProtective-serv	0.439***
occupationSales	0.241***
	(0.067)
occupationTech-support	0.515***
	(0.091)
occupationTransport-moving	-0.197**
	(0.081)
relationshipNot-in-family	-0.892***
	(0.136)
relationshipOther-relative	-1.202***
	(0.190)
relationshipOwn-child	-1.630***
	(0.168)
relationshipUnmarried	-1.176***
	(0.153)
relationshipWife	1.133***
	(0.086)
raceOther	0.887**
	(0.400)
raceWhite	0.705***
	(0.273)
genderMale	0.756***
	(0.065)
capital_loss	0.001***
	(0.00003)
capital_gain	0.0003***
	(0.00001)
hours_per_week	0.024***
	(0.001)
native_countryUnited-States	0.151**
	(0.063)
I(age2)	-0.002***
	(0.0001)
age:raceOther	-0.018**
	(0.009)
age:raceWhite	-0.011*
	(0.006)
Constant	-12.598***
	(0.401)
Observations	45,222
Log Likelihood	-14,522.530
Akaike Inf. Crit.	29,129.060
Note:	*p<0.1; **p<0.05; ***p<0.01

Results for the Logit Non-linear Model shows that all of our independent variables are statistically significant. Marginal effects are calculated using logitmfx() function from mfx package. Chosen interpretations will be presented below:

Additional year of age will increase the probability of a success by 2.5 percentage points

Being a male will increase the probability of a success by 7.1 percentage points comparing to a female

Being self-employed will decrease the probability of success by 2 percentage points comparing to being a governmental worker

Being white will increase the probability of a success by 6.0 percentage points comparing to being a black

Being graduated from college will increase the probability of a success by 0.9 percentage point comparing to those being graduated from school

Additional year of schooling will increase the probability of success by 2.9 percentage points

Being married will increase the probability of success by 10 percentage points comparing to those that are divorced

Holding an Executive Managerial position will increase the probability of success by 9.5 percentage points

Working additional hour will increase the probability of success by 0.2 percentage points

With additional year of age and being white will decrease the probability of success by 0.01 percentage point comparing to being black.

Additionally, odds ration are presented only for chosen variables.

Having a child will decrease the probability of having an income more than 50K almost twice as not having a family

Having a wife will increase the probability of having income more than 50K almost twice as being widowed

Working at Sales position will increase the probability of having income more than 50K triple times as belonging to privet work class.

Working at Prof-Specialty position will increase the probability of having income more than 50K twice as Working at Sales position

Interpretation of R-Square:

Table 3. R-Squared

R-Squared			
Tjur	McKelveyZavoina	Adjusted R^2	Count R^2
0.45	0.8	0.397	0.851

Tjur: On average having income higher than 50K is 45% higher for those who had income more than 50K in comparison to those that did not have income higher than 50K.

McKelveyZavoina: If the unobserved latent variable was observed than our model will be explained at 80% of its variation.

Adjusted count R^2: Only 40% of all observations were predicted correctly because of the variation of the dependent variable.

Count R^2: Our model correctly predicts 85% of all observations.

Next step in our analysis will be to check the significance of education and gender variable in order to understand whether these factors are important in our model.

Testing Hypothesis:

H0: education is not statistically significant

H1: education is statistically significant

Due to the small p-value which is equal to 0.00007167 we are rejecting null hypothesis stating that education is not statistically significant. Therefore we can conclude that education has impact on the level of income.

H0: gender is not statistically significant

H1: gender is statistically significant

Due to the small p-value which is equal to 0.00000000000000022 we are rejecting null hypothesis stating that education is not statistically significant. Therefore we can conclude that gender has impact on the level of income.

We can make the conclusion that gender and education do have impact on income.

To be sure that our results have sense Hosmer-Lemeshow and Linktest are conducted. Below are the results for forementioned test.

Hosmer-Lemeshow Test:

H0: Model has no omitted variables

H1: Model has omitted variables

P-value is smaller than 5% significance test hence we are rejecting null hypothesis in favor of alternative stating that our model has omitted variables.

Linktest* showed that squares of residuals are statistically significant therefore we can state that the form of our model is not correct.

Linktest*: Code for the test was prepared by Rafal Wozniak.

5. Results

It is obvious that our model has missing variables which possess important information regarding the income, so results attained from the model cannot be taken into consideration. Our model miss the information regarding performance at work, whether working field is in line with education, economic situation of countries where our respondents are based.

6. Findings

Our model miss some important information which can affect income, so we cannot take seriously results obtain in our analysis. Our assumptions about missing information are following: information regarding performance at work, whether working field is in line with education, economic situation of countries,

7. Bibliography

Duc Hong Vo , Thang Cong Nguyen , Ngoc Phu Tran and Anh The Vo, What Factors Affect Income Inequality and Economic Growth in Middle-Income Countries?

W.H. Greene, Econometric Analysis, Pearson

Examining the Factors Affecting Personal Income: An empirical Study Based on Survey Data in Chinese Cities: http://www.econ.kobe-u.ac.jp/activity/graduate/pdf/王李蕙.pdf

8. Appendix

```
library(rsample)
library(recipes)
library(tidyr)
library(stringr)
library(readr)
library(dplyr)
library(mfx)
library(DescTools)
library(LogisticDx)
library(blorr)
library(ggplot2)

options(scipen = 999)
setwd("C:\\Users\\daria\\OneDrive\\Desktop\\datasets\\data for R project")
# Data Exploration -------------------------------------------------------


# loading data
data <- read_csv('income.csv')

# dimension of the data. our dataset has 48842 rows and 15 columns.
dim(data)

colnames(data)

head(data)
```

```
# checking for missing values.
colSums(is.na(data))

# counts for different income groups
data %>% count(income_bracket)

# minimum and maximum age
min(data$age)
max(data$age)

# plots
data %>%
  mutate(income_bracket = str_replace(income_bracket, '\\.', '')) %>%
          ggplot(aes(x = age, fill = income_bracket)) + geom_histogram() + theme_bw()

data %>%
  mutate(income_bracket = str_replace(income_bracket, '\\.', '')) %>%
  ggplot(aes(x = marital_status, fill = income_bracket)) + geom_bar(position = "fill") + theme_bw() +
coord_flip()

data %>%
  mutate(income_bracket = str_replace(income_bracket, '\\.', '')) %>%
  ggplot(aes(x = gender, fill = income_bracket)) + geom_bar(position="dodge") + theme_bw()

data %>%
  mutate(income_bracket = str_replace(income_bracket, '\\.', '')) %>%
  ggplot(aes(x = race, fill = income_bracket)) + geom_bar(position="dodge") + theme_bw()+
  coord_flip()

data %>%
  mutate(income_bracket = str_replace(income_bracket, '\\.', '')) %>%
  ggplot(aes(x = income_bracket, y = hours_per_week, fill = income_bracket)) + geom_boxplot()


ggplot(data, aes(education_num)) + geom_histogram() + theme_bw()

ggplot(data = data) +
  aes(x=education_num, fill=education) +
  geom_bar() +
  theme_bw()


# counts for differnt workclass groups. We noticed "?" value in workclass variable.
```

```
data %>% count(workclass)


# fnlwgt will be excluded in the following step of data preprocessing beacause it's just a weight calculated
based on demographical
# factors already included in the dataset such as gender, nationality etc.


# Education (type and number of years)
count(data, education)

max(data$education_num)
min(data$education_num)


# counts for Martial status
count(data, marital_status)


# Data Preprocessing ----------------------------------------------------

### Cleaning the dataset ###
# Outcome variable

# Remove "." & Recode 1/0 and rename to "income"
data <- data %>%
  mutate(income_bracket = str_replace(income_bracket, '\\.', ''),
         income_bracket = ifelse(income_bracket == '<=50K', '0', '1')) %>%
  rename(income = income_bracket)

data["income"] = as.numeric(data$income)

# counts for income variable
count(data, income)

# excluding final weight (fnlwgt) from dataset

#data = select(data, -fnlwgt)

#data["fnlwgt"] = NA

data = data[-3]

# checking for missing values
```

```
colSums(is.na(data))

# replacing "?" with NA to remove observations
data[data == "?"] <- NA

data = na.omit(data)

colSums(is.na(data))

### Merge underrepresented classes ###


count(data, marital_status)

data <- data %>%
  mutate(
    marital_status = ifelse(marital_status %in% c('Married-spouse-absent',
                                                 'Married-AF-spouse',
                                                 'Married-civ-spouse'),
                            'Married', marital_status)
  )

count(data, occupation)

data <- data %>%
  mutate(
    occupation = ifelse(occupation == 'Armed-Forces', 'Protective-serv',
                        occupation)
  )


count(data, native_country) %>% arrange(-n)

data <- data %>%
  mutate(
    native_country = ifelse(native_country %in% c('United-States'),
                            native_country, 'other')
  )

count(data, education)

data <- data %>%
  mutate(
    education = ifelse(education %in% c('11th', '7th-8th', "5th-6th", '12th', "10th", "1st-4th", "9th", 'Preschool', "HS-grad"), 'School', ifelse( education %in% c("Some-college", "Prof-school"), "College", ifelse( education %in% c("Assoc-voc", 'Assoc-acdm'), "Associate", education))))



count(data, workclass)


data = data %>%
  mutate(
    workclass = ifelse(workclass %in% c("Federal-gov", "Local-gov", "State-gov"), "governmental position", ifelse(workclass %in% c("Self-emp-inc", "Self-emp-not-inc"), "Self-emp", ifelse(workclass %in% c('Without-pay', 'Never-worked'), "No-Pay", workclass)
  )))

count(data, race)
data = data %>%
  mutate(
    race = ifelse(race %in% c("Amer-Indian-Eskimo", "Asian-Pac-Islander"), "Other", race)
  )



categorical_vars <-
  sapply(data, is.character) %>%
  which() %>%
  names()

categorical_vars

for (variable in categorical_vars) {
  data[[variable]] <- as.factor(data[[variable]])
}

data = data %>%
  mutate(
    education = factor(education, levels = c('School', 'College', 'Bachelors', 'Masters', 'Doctorate', 'Associate')))

# Analysis ----------------------------------------------------------------

# Linear Probability model
lpm = lm(income ~ ., data=data)
```

```
summary(lpm)

lpm.residuals = lpm$residuals
bptest(lpm.residuals ~., data=data)

# Breusch-Pagan test shows that there is no heteroskedasticity in our data. Its a first sing showing that
there are
#some problems with our data. Following tests will provide better understandings.

# probit model
myprobit <- glm(income ~., data = data,
          family=binomial(link="probit"))
summary(myprobit)

# logit model
mylogit <- glm(income~., data = data,
          family=binomial(link="logit"))
summary(mylogit)

# Based on AIC information criteria we should choose logit model.
# All the variable in our model are statistically significant at 5% significance level.

logit_non_linear = glm(income ~ age + workclass + education + education_num+ marital_status +
occupation + relationship +
                race+ gender + capital_loss + capital_gain + hours_per_week + native_country + I(age^2) +
age*race,
              data, family=binomial(link="logit"))

summary(logit_non_linear)

#Comparing logit model with interaction and variable to power and simple logit model,
#first model will be chosen based on AIC information criteria

# creating quality table
library(stargazer)
stargazer(myprobit, mylogit, logit_non_linear, lpm, type="text")

# Adds ratio
1.63 / 0.88
#Having a child will decrease the probability of having an income more than 50K almost twice as not having
a family
1.12 / 0.57
# Having a wife will increase the probability of having income more than 50K almost twice as being
widowed
```

```
0.24 / 0.08
# Working at Sales position will increase the probability of having income more than 50K triple times as belonging to privet workclass
0.47 / 0.24
# Working at Prof-Specialty position will increase the probability of having income more than 50K twice as Working at Sales position

# Marginal effects

marr_effects=logitmfx( formula = income ~ age + workclass + education + education_num+ marital_status + occupation + relationship +
          race+ gender + capital_loss + capital_gain + hours_per_week + native_country + I(age^2) + age*race,
        data, atmean = T)

class(marr_effects)



# Interpretation of R^2
PseudoR2(logit_non_linear, c=( 'Tjur', 'McKelveyZavoina'))


# adjusted R^2
blr_rsq_adj_count(logit_non_linear) # function from package blorr

#count R^2
blr_rsq_count(logit_non_linear) # function from package blorr


# Testing hypothesis

# H0: education = 0
# H1: education != 0

logit_unrestricted = glm(income ~ age + workclass + race + education + education_num+ marital_status + occupation + relationship +
                    race+ gender + capital_loss + capital_gain + hours_per_week + native_country + I(age^2) + age*race,
                   data, family=binomial(link="logit"))

# removing education variable from the model
logit_restricted = glm(income ~ age + workclass + race + education_num+ marital_status + occupation + relationship +
          race+ gender + capital_loss + capital_gain + hours_per_week + native_country + I(age^2) +
age*race,
        data, family=binomial(link="logit"))

lrtest(logit_unrestricted, logit_restricted)


# H0: gender = 0
# H1: gender != 0

logit_unrestricted1 = glm(income ~ age + workclass + race + education + education_num+ marital_status
+ occupation + relationship +
                   race+ gender + capital_loss + capital_gain + hours_per_week + native_country + I(age^2) +
age*race,
                 data, family=binomial(link="logit"))

# removing education variable from the model
logit_restricted1 = glm(income ~ age + workclass + race + education + education_num+ marital_status +
occupation + relationship +
                 race + capital_loss + capital_gain + hours_per_week + native_country + I(age^2) + age*race,
                data, family=binomial(link="logit"))

lrtest(logit_unrestricted1, logit_restricted1)


# Hosmer-Lemeshow test

# H0: Model has no omitted variables
# H1: Model has omitted variables

gof.results = gof(logit_non_linear)
gof.results$gof


# Link Test

source("linktest.R")
linktest_results = linktest(logit_non_linear)
summary(linktest_results)

#quality table
stargazer(logit_non_linear, type = "text",  title="Logit non-linear Results", out = "table1.txt")
```